Improving the Accuracy of Multinomial Naïve-Bayes Algorithm with Adaptive Boosting Using Information Gain for Classification of Movie Reviews Sentiment Analysis

Hanif Nur Cahyani [1,*], Riza Arifudin [1]

[1] Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: hanifnurcahyani19@gmail.com

ARTICLE INFO

Article history
Received 9 Maret 2022
Revised 25 Maret 2022
Accepted 16 April 2022

Keywords
IMDb
Sentiment Analysis
Multinomial Naïve-Bayes
Adaptive Boosting
Information Gain

ABSTRACT

Movie is a means of delivering information as well as entertainment that can be enjoyed by all people through various platforms such as the internet, cinema, and television. Sentiment analysis is needed to analyze positive and negative comments from movie lovers, these comments come from many circles and from various sources, one of which is IMDb (Internet Movie Database). Naïve Bayes multinomial classification algorithm has been proposed and used by many researchers in the case of sentiment analysis. An ensemble Adaptive Boosting algorithm is used as a boosting algorithm to improve accuracy in Naïve Bayes and multinomial classification models of information acquisition. The accuracy test on the model is carried out using the python programming language. The accuracy results obtained when applying the Naïve Bayes multinomial classification algorithm is 84.82%, then an accuracy of 85.24% is obtained when implementing information gain feature selection in the Naïve Bayes multinomial classification algorithm. The highest accuracy result of 87.87% was obtained when implementing the multinomial Naïve Bayes classification algorithm with Adaptive Boosting and Information Gain Selection features.

This is an open access article under the CC–BY-SA license.

1 Introduction

Movie or film is a means of conveying various information to the public through story media. Film can also be interpreted as a medium of artistic expression for artists and filmmakers to express their ideas and story ideas (Ratnawati, 2018). The quality of the film can determine the success of the film in the eyes of the audience. The quality of the film itself can be concluded based on comments, both positive and negative comments from the audience (Sudiantoro et al. 2018). Sentiment analysis is a process of understanding, extracting and processing textual data automatically to obtain sentiment information contained in a sentence, such as in movie user reviews for example (Sudiantoro et al. 2018). The definition of classification according to Bunga et al. (2018) is a function that is predictive by entering into certain data groups into classes. Classification is made based on a set of training sets with classes that have been determined based on their characteristics. Classification can predict a class in a document (Aliwy & Ameer, 2017). Classification will assign the class correctly into a new document that is obtained from a collection of several classes (Hamzah, 2021).

The multinomial Naïve Bayes algorithm is an algorithm has been used by Kalcheva et al. (2020) in the case of a classification analysis of Bulgarian literature. According to Kalcheva et al. (2020) Multinomial Naïve Bayes algorithm is one of the more accurate classification algorithms with faster computation time. From several studies have been carried out by several researchers above, it

can be obtained information that the multinomial Naïve Bayes algorithm is one of the most widely used classification algorithms and it is suitable to be implemented in text mining research. In the use of the Naïve Bayes multinomial algorithm, of course, it can be optimized with other algorithms. One of the ensemble methods that can optimize the multinomial Naïve Bayes algorithm is Adaptive Boosting or can also be called Adaboost (Hamzah, 2021).

Adaboost is one of the ensemble methods that is often used in various studies works by approaching the Naïve Bayes classifier by combining several weak classifiers or weak learners, then producing a single strong classifier for multiclass classification problems (Vergara et al. 2016). One of the problems are often encountered in implementing the multinomial Naïve Bayes algorithm is there are still excessive irrelevant or unnecessary features and this can reduce the performance of the algorithm itself (Xue et al. 2014). These irrelevant features can be in the form of words that have less correlation with the object of research. Features that do not have a correlation with the object of research are not really needed when classifying data. Therefore, it is necessary to implement a feature selection algorithm to handle these problems.

The selection feature implemented in this study is information gain which is used to optimize and improve the results of data testing accuracy on the multinomial Naïve Bayes algorithm with Adaboost. Information gain as a feature selection algorithm has been used by Somantri & Apriliani (2018) in their research to compare the level of accuracy resulting from the application of Chi-Square Statistical feature selection and information gain on the Support Vector Machine (SVM) algorithm. In their research, Somantri & Apriliani (2018) concluded that the Information Gain Selection feature has better performance with a higher level of accuracy than the Chi-Square Statistical feature.

Based on the description of the problem above, this research focuses on increasing the accuracy of the multinomial Naïve Bayes algorithm with Adaboost using the Information Gain Selection feature in the classification of movie reviews sentiment analysis

2 Method

This research using multinomial Naïve Bayes algorithm with Adaboost to optimize the result and want to increase in accuracy using the information gain selection feature which was chosen because it to reduce irrelevant features and the dimensions of features in the data (Sari, 2016). The schematic of the proposed model flow can be seen in Figure 1.

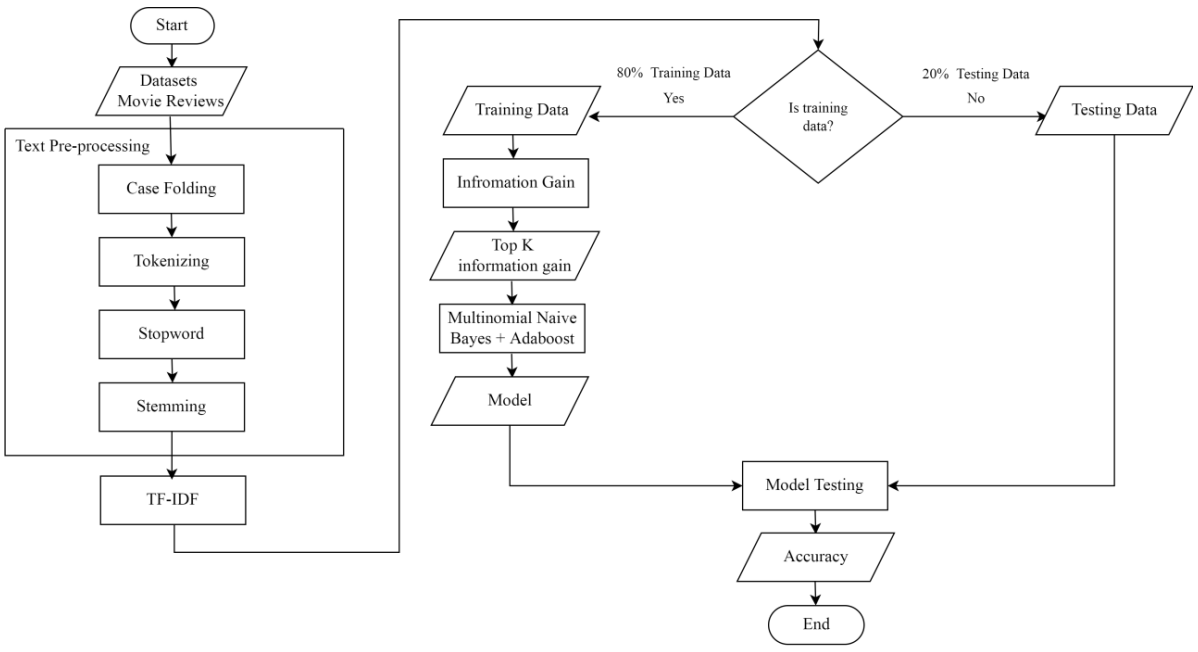

Figure 1. Flowchart of the proposed method

After the dataset is collected, the next step in this research is to perform the preprocessing stage of the data. The preprocessing result data is then converted into a number vector using Term Frequency – Inverse Document Frequency (TF-IDF). TF-IDF data results then calculated the weight of the highest gain value from the Information Gain with the best top k value in the information gain classification process with Naïve Bayes and Adaboost. The last process is testing the model of the applied methods so accuracy results are obtained.

2.1 Data Collection

In this study, the researcher used data as the object of his research. The data used in this study is a collection of data taken from the dataset of the movie review dataset on the large movie review v1.0 dataset. Which is accessed through the https://ai.stanford.edu/~amaas/data/sentiment/ page on January 10, 2022. The dataset consists of 50,000 film review data containing two attributes, namely sentiment attribute and review attribute, with a total of 50,000 positive sentiments. 25,000 documents and 25,000 negative sentiment documents. This study uses 50,000 data, then the data is divided into training data and testing data, the comparison for training data with testing data is 80:20, as much as 80% data for training data, and as much as 20% data for testing data.

2.2 Preprocessing

The dataset obtained from the large movie review v1.0 dataset, then carried out the data preprocessing stage by going through four stages, it is the Case Folding, Tokenizing, Stopword and Stemming stages as described below.

- Case Folding is a process used to equalize the shape of the letters, for example from capital letters to all non-capital letters or vice versa (Kowsari et al. 2019).
- Tokenizing is a popular technique for lexican analysis designed to break down a document or text into individual words (Susilowati et al. 2015).
- Stopword is a list of words that have no effect in a document (Handayani & Pribadi, 2015).
- Stemming is the process of simplifying a word into a basic word (Ipmawati et al. 2017).

2.3 Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) is a feature extraction algorithm. Dey et al. (2016) reveal the working mechanism of Feature Extraction is to calculate numerical values or information symbolically from an observation. The importance of characteristic words/terms in text concentration will increase with increasing word frequency in each document, but will be inversely proportional to word frequency in all text concentrations (Zhu et al. 2019).

Saadah et al. (2013) wrote the TF-IDF equation as in Equation 1 to calculate the term frequency (tf) with the i term frequency ($tf_{(i)}$) being the number of occurrences of the i term ($freq_i$) in the j document. (d_j).

$$tf(i) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_{ij})} \quad (1)$$

Equation 2 is a calculation for the i inverse document frequency (idf_i) is the logarithm of the ratio of the total number of documents (D) in the corpus according to the number of documents that have terms.

$$idf_i = log\frac{|D|}{|\{d : t_i \in d\}|} \quad (2)$$

then the value will be obtained in Equation 3, namely by multiply both.

$$tfidf = tf + idf \quad (3)$$

2.4 Splitting Data

Using a total dataset of 50,000 documents, the data will be split with a ratio of 80% and 20%. A total of 40,000 data will be used as training data and 10,000 data will be used as testing data. The

reason for implementing 80:20 data splitting is based on the Pareto principle. Pareto got the result that 20% of factors will produce 80% of other factors. In addition, the 80:20 comparison was chosen because the comparison is most commonly used for split data with good performance.

2.5 Information Gain Feature

Information Gain is a feature selection method or feature selection that is commonly used in sentiment analysis. The workings of feature selection are based on large feature space reductions, namely by eliminating attributes that are less relevant and by using the appropriate feature selection algorithm so as to increase accuracy (Jindal et al. 2015). Information Gain Selection feature algorithm is one of the best feature selection algorithms that is often used in classifying text data. For the calculation of the entropy value on the Information Gain Selection feature, can be seen in Figure 2.

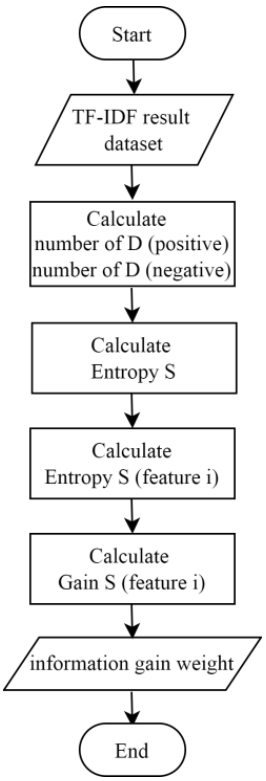

Figure 2. Flowchart of the information gain

2.6 Multinomial Naïve-Bayes

Naïve-Bayes multinomial is a development of the Naïve Bayes algorithm which is based on the Naïve Bayes theorem. The way of Naïve Bayes algorithm works is to predict future probabilities based on previous experience (Hamzah, 2021). The theorem of naïve Bayes multinomial equations can be seen in equation 4 and equation 5.

$$\hat{P}(c) = \frac{N_c}{N} \tag{4}$$

$$\hat{P}(c) = \frac{T_{ct}+\alpha}{(\sum_{t' \in V} T_{ct'})B'} \tag{5}$$

$\hat{P}(c)$ = Prior probability class c

N_c = Number of document class c

N = Total number of documents

T_{ct} = Number of term or word t in documents with category c

$\sum_{t' \in V} T_{ct'}$ = Total number of frequency term class c

α = Value of Laplacian smoothing betwen $0 < \alpha \leq 1$

B' = Total number of term vocabulary

2.7 Adaptive Boosting

Boosting is an ensemble method that works sequentially where most of the base learners consist of the same type homogenous (Hamzah, 2021). Boosting works by correcting errors contained in each previous iteration, so the resulting base learner will give more focus to patterns that may be difficult to classify. Boosting is widely used because it is considered effective in terms of classification problems by changing the weights on the sample training data (Tan et al. 2019). According to Vergara et al. (2016) the Adaboost algorithm is an iterative procedure that approaches the Bayes classifier by combining several weak classifiers or weak learners, then producing a single strong classifier for multiclass classification problems. From some of these statements it can be said that Adaboost is a boosting algorithm that is used to combine the output of weak learners that works by reducing the next weakness that is misdiagnosed in the weak learner on the previous misclassification record by weighting each learner to get the final output. Adaboost will minimize errors in each iteration at the time of weighting. This algorithm is sensitive to noise and relatively less prone to overfitting.

3 Results and Discussion

This research discusses about sentiment analysis of the movie reviews v1.0 dataset, with accuracy testing through a combination of application of classification algorithms, selection features and ensembles. The classification algorithm used is the multinomial Naïve Bayes algorithm, with the ensemble Adaboost method and Information Gain Selection feature. This study compares the increase in accuracy results with other sentiment analysis studies but with different methods in each study.

The first process that is carried out on the dataset is preprocessing the data. The results of the preprocessing data can be seen in Table 1.

Table 1. Text Preprocessing result

Reviews	Method
this a fantastic movie of three prisoners who become famous one of the actors is george clooney and im not a fan but this roll is not bad another good thing about the movie is the soundtrack the man of constant sorrow i recommand this movie to everybody greetings bart	Case folding
thi fantast movi three prison becom famou one actor georg clooney im fan thi roll bad anoth good thing movi soundtrack man constant sorrow recommand thi movi everybodi greet bart	Tokenize
thi fantast movi three prison becom famou one actor georg clooney im fan thi roll bad anoth good thing movi soundtrack man constant sorrow recommand thi movi everybodi greet bart	Stopword
thi fantast movi three prison who becom famou one actor georg clooney and im not fan but thi roll not bad anoth good thing about the movi soundtrack man constant sorrow recommand movi	Stemming

The next process is to change the word vector in the text document into a numeric vector using the TF-IDF word vectorization method. The process in the TF-IDF method produces a weight value for each term/word in each document obtained from the calculation of the value of the number of occurrences of the term/word in each document (TF) and the calculation of the inverse frequency value of the document (IDF), then from the TF value and IDF value will be used to obtain the result of the TF-IDF value to get numeric vector result. The results of the TF-IDF process can be seen in Table 2.

Table 2. Numeric vector result

Term	Numeric Vector
review	1.43136376416
fact	1.43136376416
give	0.26413688858
never	0.26413688858
wonder	0.26413688858
sens	0.26413688858
actor	1.43136376416
dare	0.26413688858
perform	1.43136376416
plot	1.43136376416
done	1.43136376416
come	0.26413688858
kill	1.43136376416

The next step after getting the number vector is testing the top k value on the Information Gain which is tested with the multinomial Naïve Bayes classification method. the graph for the results of the best top k test, the information gain can be seen in Figure 3.

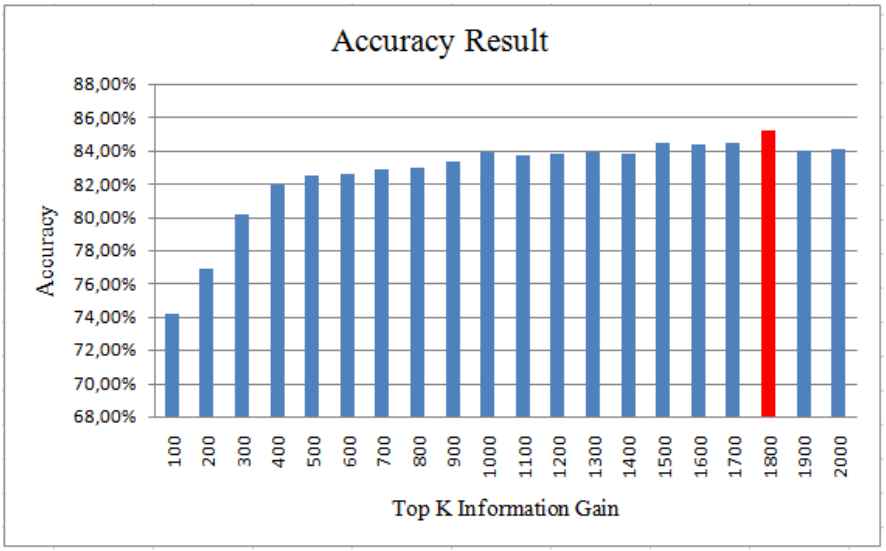

Figure 3. Accuracy from the experiment top k information gain

Based on the top k information gain test, the highest accuracy was obtained with 85.24% results at the top k 1.800. This shows that the best features obtained in the Information Gain process are 1.800 word features. The application of Information Gain in this study was able to increase the accuracy from 84.28% when using multinomial Naïve Bayes alone to 85.24%. The next process is testing the Adaboost parameters. The results of testing the adaboost parameters can be seen in Table 3.

Table 3. Adaboost parameter test results

Learning Rate	Estimator				
	50	75	100	200	**400**
0.1	75.18	76.78	78.47	81.10	83.67
0.3	80.38	82.19	83.38	84.89	86.9
0.5	82.87	83.42	84.00	85.94	**87.87**
0.8	82.25	83.43	84.11	86.11	87.67
1.0	81.69	8341	83.64	85.67	86.83

At the stage of applying the multinomial Naïve Bayes algorithm classification with Adaboost and Information Gain Selection feature, several Adaboost parameters will be tested, namely the estimator and learning rate. Several numbers of Adaboost parameters were tested in the research, the results were the learning rate or iteration and the estimator or number used to weight each iteration which was the most optimal when the learning rate value was 0.5 and the estimator was 400 with an accuracy of 87.87%. To prove that model testing can increase accuracy, Figure 4 is a graph of increasing accuracy of the model that has been tested in the study.

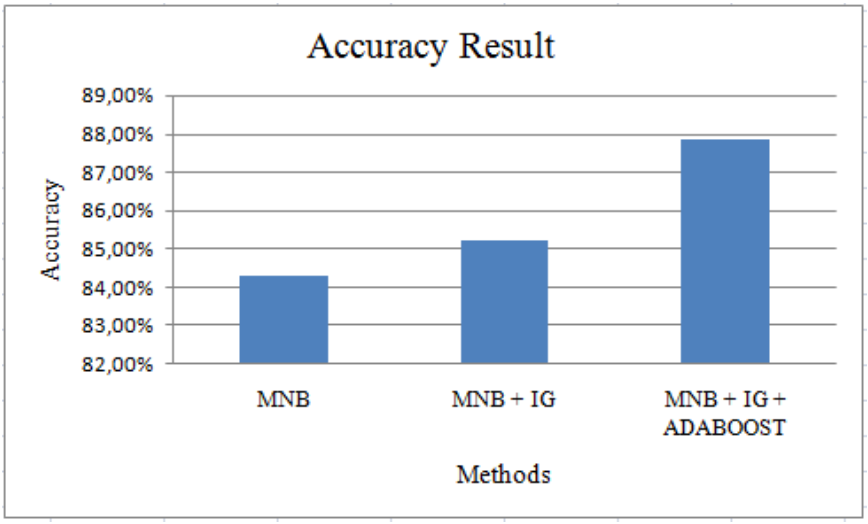

Figure 4. The results of the accuracy of the proposed method

In order to know whether the method that has been applied in this study can be said to be better than the method used in previous studies, this study compares the results of model testing accuracy with several previous studies using the same dataset or with different datasets. The results of the comparison of accuracy with previous studies can be seen in Table 4.

Table 4. Comparison of accuracy with previous research

Writer	Method	Accuracy
Baid et al., (2017)	Naïve Bayes	81.4%
	K-Nearest Neighbour	55.30%
	Random Forest	78.65%
Somantri, & Apriliani (2018)	Support Vector Machine	69.36%
	Support Vector Machine + Information Gain	72.45%
	Support Vector Machine + Chi Squared Statistic	70.09%
Kalcheva et al., (2020)	Multinomial Naïve Bayes	80.0%
	Bernoullinaïve Bayes	64.0%
	Support Vector Classification	88.7%
	Random Forest	74.7%
	Adaboost	82.0%
Hamzah (2021)	Multinomial Naïve Bayes	81.39%
	Multinomial Naïve Bayes+ Chi Squared Statistic	85.37%
	Multinomial Naïve Bayes+ Chi Squared Statistic + Adaboost	87.74%
Proposed Method	Multinomial Naïve Bayes	84.28%
	Multinomial Naïve Bayes + Information Gain	85.24%
	Multinomial Naïve Bayes + Adaboost + Information Gain	87.87%

This study was able to produce the highest accuracy after the application of the multinomial Naïve Bayes algorithm with Adaboost and Information Gain on the movie reviews v1.0 dataset of 87.87% on accuracy testing with the Adaboost parameter tuning with an estimator value of 400 and a learning rate of 0.5. This is slightly different from the previous research conducted by Hamzah (2021) when applying the multinomial Naïve Bayes algorithm with Adaboost and Chi Square Statistics as a feature selection algorithm for accuracy testing with an estimator value of 100 and a learning rate of 0.1 Adaboost which produces an accuracy of 87,74%.

The accuracy results obtained from this study can be said to be better than the research that has been done by previous researchers in the case of sentiment analysis classification using the movie reviews v1.0 dataset. The application of the Information Gain selection feature with a top k value of 1.800 in this study was able to increase the accuracy of the multinomial Naïve Bayes classification algorithm with Adaboost. In this study, it can produce the best accuracy rate of 87.87% when using Adaboost at an estimator value of 400 and a learning rate of 0.5 where the estimator and learning rate values are still said to be too large and take too long.

4 Conclusion

The application of Information Gain in this study is functioned as a selection feature that works by selecting features that are less relevant based on the results obtained from the gain calculation for each feature. The feature with the lowest gain value will then be selected by Information Gain, a feature with a high gain value will be obtained which is considered the most relevant to the classification. With the application of feature selection on the multinomial Naïve Bayes classification algorithm with Adaboost, this study obtained better accuracy results than the accuracy test without the application of feature selection. The accuracy results obtained from this study after the application of the Information Gain and Adaboost selection features on the Naïve Bayes multinomial classification algorithm are 87.87%, the accuracy results have increased by 3.59% compared to those using only the multinomial Naïve Bayes classification algorithm, or an increase of 2.63% compared to when using multinomial Naïve Bayes and Information Gain.

References

Aliwy, A. H., & Ameer, E. H. A. (2017). Comparative study of five text classification algorithms with their improvements. *International Journal of Applied Engineering Research*, *12*(14), 4309–4319.

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, *179*(7), 45-49.

Bunga, M. T. H., S, B., Djahi, & Nabuasa, Y. Y. (2018). Multinomial Naive Bayes Untuk Klasifikasi Status Kredit Mitra Binaan Di Pt . Angkasa Pura I Program Kemitraan. *J-Icon*, *6*(2),30–34.

Dey, S., Kumar, Y., Saha, S., & Basak, S. (2016). Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting. *PESIT South Campus.*, 1–10.

Hamzah, M. B. (2021). Classification of Movie Review Sentiment Analysis Using Chi-Square and Multinomial Naïve Bayes with Adaptive Boosting. *Journal of Advances in Information Systems and Technology*, *3*(1), 67–74.

Handayani, F., & Pribadi, S. (2015). Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110. *Jurnal Teknik Elektro*, *7*(1), 19–24.

Ipmawati, J., Kusrini, & Taufiq Luthfi, E. (2017). Komparasi Teknik Klasifikasi Teks Mining Pada Analisis Sentimen. *Indonesian Journal on Networking and Security*, *6*(1), 28–36.

Jindal, R., Malhotra, R., & Jain, A. (2015). Techniques for text classification: Literature review and current trends. *Webology*, *12*(2), 1–28.

Kalcheva, N., Karova, M., & Penev, I. (2020). Comparison of the accuracy and the execution time of classification algorithms for Bulgarian literary works. *2020 International Conference Automatics and Informatics, ICAI 2020 - Proceedings*, *1*.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information (Switzerland)*, *10*(4), 1–68.

Ratnawati, F. (2018). Implementasi Algoritma Naive Bayes Terhadap Analisis Sentimen Opini Film Pada Twitter. *INOVTEK Polbeng - Seri Informatika*, *3*(1), 50.

Saadah, M. N., Atmagi, R. W., Rahayu, D. S., & Arifin, A. Z. (2013). Sistem Temu Kembali Dokumen Teks Dengan Pembobotan Tf-Idf Dan Lcs. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *11*(1), 19.

Sari, B. N. (2016). Implementasi Teknik Seleksi Fitur Information Gain Pada Algoritma Klasifikasi Machine Learning Untuk Prediksi Performa Akademik Siswa. *Seminar Nasional Teknologi Informasi Dan Multimedia 2016*, *March*, 55–60.

Somantri, O., & Apriliani, D. (2018). Support Vector Machine Berbasis Feature Selection Untuk Sentiment Analysis Kepuasan Pelanggan Terhadap Pelayanan Warung dan Restoran Kuliner Kota Tegal. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(5), 537.

Sudiantoro, A. V., Zuliarso, E., Studi, P., Informatika, T., Informasi, F. T., Stikubank, U., & Mining, T. (2018). Analisis Sentimen Twitter Menggunakan Text Mining Dengan Algoritma Naive Bayes Classifier. *Dinamika Informatika*, *10*(2), 398–401.

Susilowati, E., Sabariah, M. K., & Gozali, A. A. (2015). Implementasi Metode Support Vector Machine untuk Melakukan Klasifikasi Kemacetan Lalu Lintas Pada Twitter. *E-Proceeding of Engineering*, *2*(1), 1478–1484.

Tan, Z., Zhang, Y., Zhang, C., Huang, R., Lei, P., & Duan, X. (2019). Research on the Text Emotion of Multinomial Naïve Bayes Integration Algorithm. *Proceedings of 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC 2019*, *Imcec*, 107–111.

Vergara, D., Hernandez, S., & Jorquera, F. (2016). Multinomial Naive Bayes for real-time gender recognition. *2016 21st Symposium on Signal Processing, Images and Artificial Vision, STSIVA 2016*, 2–7.

Xue, B., Zhang, M., & Browne, W. N. (2014). Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. *Applied Soft Computing Journal*, *18*, 261–276.

Zhu, Z., Liang, J., Li, D., Yu, H., & Liu, G. (2019). Hot Topic Detection Based on a Refined TF-IDF Algorithm. *IEEE Access*, *7*(c), 26996–27007.